LLM Evaluation in Practice: A Review of Metrics, Practitioner Insights, and Lessons Learned

R.M. Bakker[1,2], M.W. Witte-Schaaphok[1], J. García-Fernández[1], T. Brand[1], J. van der Weide[1], S.A. Raaijmakers[1,2]

[1] The Netherlands Organization for Applied Scientific Research (TNO),
[2] Leiden University Centre for Linguistics
{roos.bakker, marianne.schaaphok, julia.garciafernandez, tom.brand, jens.vanderweide,stephan.raaijmakers}@tno.nl

Abstract

The rapid, widespread adoption of Large Language Models (LLMs) highlights the need to understand their performance, strengths, and limitations. However, evaluating LLMs presents significant challenges due to the broad range of tasks and model capabilities, especially in practice or low-resource settings where benchmark datasets are not available. In text generation tasks, answer diversity has always complicated automatic evaluation, and the enhanced fluency and creativity of LLMs lead to further challenges. Existing metrics and frameworks often fail to account for these complexities. Furthermore, recent research into the replicability of benchmarks has demonstrated serious issues when reproducing historical benchmark results. This paper makes two key contributions: (1) a categorisation of challenges and metrics in LLM evaluation, and (2) lessons learned from practice through a survey and a use case. To this end, a literature study was conducted to identify challenges and metrics in scientific work. A survey among developers working with LLMs provided insights into practical challenges. Furthermore, selected metrics were implemented in a practical use case to gain insights into their strengths and limitations. By combining theoretical analysis with real-world experiences and lessons learned from practice, this work provides an overview and best practices for users evaluating LLM performance.

Keywords: LLM Evaluation, NLP Evaluation, Applied Evaluation

1. Introduction

The evaluation of generative Large Language Models (LLMs) has become increasingly critical as these models have gained widespread use in recent years. In the past, language models were evaluated on specific tasks using manual annotations against benchmarks. While this approach remains effective for classification tasks, it is not always suitable for text generation due to the variability of natural language. Correct answers can be formulated in many acceptable ways, which makes it difficult to compare LLM answers to a fixed ground truth. Another challenge is that commonly used benchmarks are not always representative of real-world applications or practical usage scenarios (Kiela et al., 2021). Furthermore, many benchmarks are publicly available and may be included in LLM training (Ravaut et al., 2024), raising concerns about overfitting (Zhang et al., 2024). In this work, we contribute to a better understanding of these challenges in a practical setting through two contributions: (1) a categorisation of the key challenges and metrics in evaluating LLMs, (2) lessons learned from practice through a survey and a use case.

This paper is structured as follows: in Section 2, we will discuss related work on LLMs, LLM evaluation, and we will present a categorisation of evaluation challenges. In Section 3, we will compare a set of evaluation metrics. In Section 4, we will discuss evaluation challenges in practical contexts, gathered through a survey and a practical use case. In Section 5, we will discuss the insights from the survey and the use case. Finally, we will conclude our work and discuss ideas for future work.

2. Related Work

LLMs are a progression of standard statistical language models that condition word generation on word context by assigning probabilities to word sequences. LLMs are produced by deep neural networks based on Transformer architectures (Vaswani et al., 2017), and condition the generation of words on vectorised representations of left context. In the original Transformer architecture, a separate bi-directional encoder optimises this vectorisation in conjunction with a left-to-right operating, autoregressive decoder that provides feedback to the encoder. Most contemporary LLMs, such as GPT-3, are decoder-only architectures that integrate the encoding process within the model itself (Brown et al., 2020). These models are large according to three dimensions: the amount of words they are initially trained on, the amount of parameters in the underlying neural network models, and the training budget (Minaee et al., 2025; Naveed et al., 2024).

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 23–38
11 May 2026.

2.1. LLM Evaluation

LLMs rapidly gained popularity across diverse applications, increasing the need for robust evaluation. Traditional evaluation methods, such as perplexity, BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004), have been widely used. However, due to the fast advancements of LLMs, those metrics no longer capture the full scope of model performance (Meister and Cotterell, 2021; Wu et al., 2023a). This sparked a search for new metrics and a deeper understanding of existing ones.

Several surveys review evaluation methods for text generation and large language models, and try to categorise methods along different dimensions. For instance, the surveys of Celikyilmaz et al. (2021); Belz and Reiter (2006) make a distinction between intrinsic evaluation, which focuses on model behaviour, and extrinsic evaluation, which considers downstream task performance. The recent survey of Chang et al. (2024) takes a different approach with grouping tasks ("what"), datasets ("where"), and protocols ("how") (Chang et al., 2024) or automatic versus human (Belz and Reiter, 2006). Some surveys take a more narrow scope, such as benchmarks only (Ni et al., 2025) or evaluation challenges and limitations (Laskar et al., 2024) (which we further discuss in Section 2.1. Further categories can be added, such as alignment and safety (Guo et al., 2023; Liu et al., 2023b), but it is not always clear how those align with existing work, reflecting the lack of consensus on evaluation standards (Chang et al., 2024).

We introduce a categorisation of aspects of LLM evaluation in Table 1, which includes categories such as the ones discussed above. At the top level, we have four categories. The scope indicates the focus of the evaluation, which can be on intrinsic qualities, such as correctness of generated texts (Celikyilmaz et al., 2021), or extrinsic qualities such as a user's improved comprehension after interacting with a downstream task (Belz and Reiter, 2006). The second category of evaluation is by the approach or method of evaluation: by humans, metrics, benchmarks, or a hybrid form. Metrics and benchmarks can both be considered automatic methods (Celikyilmaz et al., 2021), align with the "where" category of Chang et al. (2024), and fall under the 'organisation' category of Guo et al. (2023). Third, evaluation can be distinguished by task type, similar to Chang et al. (2024). Within the task types, a broad distinction can be made between Natural Language Understanding (NLU) and Natural Language Generation (NLG) (Jurafsky and Martin, 2008; Khurana et al., 2023). NLU tasks, like sentiment analysis or semantic role labelling, are generally easier to evaluate using accuracy or F1. NLG tasks, such as summarisation or question answering, are more challenging due to the diversity of valid outputs and the need to assess correctness, fluency, and relevance (Liu et al., 2023a). Finally, we identify four levels of evaluation: input or prompt level, output, model, and system level.

LLM Evaluation			
Scope	**Approach**	**Task**	**Level**
Intrinsic	Human	NLG	Input
Extrinsic	Metrics	NLU	Output
	Benchmarks	Combined	Model
	Hybrid		System

Table 1: Categorisation of LLM evaluation dimensions.

2.2. Evaluation Challenges

Several challenges recur in the literature, summarised in Table 2 across scope, approach, task, and level. Many challenges cut across multiple evaluation aspects; for example, sustainability concerns can arise at multiple tasks and approaches. We focus on representative challenges at the scope and approach levels to illustrate key issues, rather than attempting an exhaustive list.

Intrinsic Challenges Intrinsic challenges can be summarised by a lack of robustness, reproducibility, variation in possible answers and a lack of model transparency. The lack of robustness can be explained by the fact that LLMs have an inherent instability, which can result in different answers for small variations in prompts. Therefore, different punctuation, wording, or order of prompts can affect the outcome of the system (Sclar et al., 2024) (Loya et al., 2023). Mizrahi et al. (2024) and Hida et al. (2024) show that varying the prompts and the prompt formatting can result in different rankings of LLMs in terms of performance. Additionally, for few-shot learning, the order of samples affects performance (Lu et al., 2022), and similarly, the order of answers for multiple choice tasks also matters (Pezeshkpour and Hruschka, 2024).

Related to this is the lack of reproducibility: generally, LLMs lack the ability to consistently obtain the same results under the same conditions due to their non-deterministic nature and confounding variables such as errors in benchmarks and different LLM model versions. This affects evaluation trustworthiness on local (where the same task can yield inconsistent results) and global scale (where reproducing research results is hindered by insufficient documentation and version control) (Laskar et al., 2024; Biderman et al., 2024). We note that non-determinism can be limited through greedy decoding settings, though this is not common practice. Recent research by Vaugrante et al. (2024) demonstrates that many benchmark results that led

Challenge	Scope	Approach	Task	Level
Robustness	Intrinsic	Any	NLG	Model
Reproducibility	Intrinsic	Any	NLG	Model
Model transparency	Intrinsic	Any	NLG	Model
Lack of standardisation	Extrinsic	Human	Any	Any
Subjectivity	Extrinsic	Human	Any	Any
Data leakage	Extrinsic	Metrics	NLU	System
Data quality	Extrinsic	Benchmarks/Metrics	Any	Output
Sustainability	Extrinsic	Any	Any	System
Bias and fairness	Intrinsic/Extrinsic	Any	Any	Any
Answer diversity	Intrinsic/Extrinsic	Metrics	NLG	Output

Table 2: LLM evaluation challenges mapped to scope, approach, task, and level.

to published landmark results (such as the effectiveness of zero-shot chain-of-thought prompting, expert prompting and sandbagging) are not reproducible, even when the same benchmarks are run again on the same model versions. This is indicative of a fundamental replicability "crisis" in LLM research. Further, and importantly, the statistical underpinning of evaluation results is only a relatively new topic in the LLM field (see e.g. (Miller, 2024)).

Sallou et al. (2024) identify three key reproducibility challenges: output variability, time-based output drift (due to retraining or user feedback), and traceability (linking outputs to specific prompts and configurations). Additionally, Atil et al. (2024) note that LLM stability varies across tasks and is rarely deterministic.

Another challenge that can be both intrinsic and extrinsic is the diversity of correct answers in NLG tasks (Wang et al., 2023). Lexical and semantic matching techniques can evaluate answers to a certain degree, but it is not useful if the given answer is not included in the ground truth (Kamalloo et al., 2023). Finally, lack of model transparency complicates LLM evaluation. Many models do not disclose their training data, architecture, or weights, making them harder to assess (Liu et al., 2023c; Liesenfeld and Dingemanse, 2024). Even with disclosed data, the low explainability of LLMs limits understanding of how they produce answers (Wu et al., 2023b)

Human-related Evaluation Challenges We can identify two evaluation challenges that mainly relate to human aspects: 1) lack of universally accepted standards, and 2) human subjectivity in evaluation. The rapid growth of LLMs has led to numerous benchmarks and evaluation methods, but no universally accepted standard exists. Researchers create their own benchmarks, raising challenges in benchmark selection, implementation, prompt variations, and fair model comparison (McIntosh et al., 2024; Post, 2018; Biderman et al., 2024). While human evaluation is essential, it remains challenging due to inter-annotator disagreement, biases, and sensitivity to question framing (Abeysinghe and Circi, 2024). Besides, human evaluation is often time-consuming and costly. Subjectivity also affects automatic evaluation, as benchmarks, examples, and annotations are influenced by human judgment.

Metrics & Benchmark Challenges There are several challenges due to the limitations of current evaluation metrics and benchmarks. First of all, data leakage or data contamination gives rise to the question whether results on benchmarks and test sets can still be trusted (Sainz et al., 2023; Zhou et al., 2023). Balloccu et al. (2024) show that the GPT-3.5 and GPT-4 models have been exposed to 4.7M samples from 263 different benchmarks during retraining. Second, whether models treat individuals or social groups fairly has been a complex, yet pressing evaluation challenge. Even current state-of-the-art LLMs have been proven to exhibit biased behaviour (Plaza-del Arco et al., 2024). There is no consensus on the conceptualisation of 'social bias' in its many forms (Blodgett et al., 2020). For LLMs specifically, existing benchmarks have been shown to be inconsistent in measuring different forms of bias (Blodgett et al., 2021). While there exist many different metrics and datasets, a 'gold standard' is lacking (Gallegos et al., 2024).

Furthermore, there is the issue of data quality. A ground truth dataset is often required, but they are costly to construct and thus limited in size and variety (Nasution and Onan, 2024). Annotation by humans is not always consistent (Hashemi et al., 2024), and annotator quality also influences the quality of the dataset, and therefore the evaluation (Grosman et al., 2020; de León Languré and Zareei, 2024). Finally, LLMs require a large amount of computing resources. A challenge is how to measure the sustainability of both training and using these models (Khowaja et al., 2024).

3. Metrics

When evaluating LLMs in practice, relying on human evaluation is often costly and time-intensive, while standard benchmarks may be unsuitable because they are neither domain- nor data-independent. As a result, automatic metrics are frequently used as a practical alternative. This section presents an overview of existing metrics, their strengths, and weaknesses.

In general, the assessed metrics show moderate correlation with human judgment and are inadequate for evaluations requiring 100% correlation with human evaluators due to the subjective nature of text evaluation (Celikyilmaz et al., 2021; Zhu et al., 2024; Liu et al., 2023a). Most metrics require references (a ground truth), which are often low-quality or unavailable (Ke et al., 2022). Reference-free metrics need unambiguous criteria definitions. Metrics like BLEU (Papineni et al., 2002) can vary in implementation, affecting comparability across studies (Post, 2018). There is no ultimate metric for LLM evaluation; the choice depends on the task, scope, and level of evaluation, the level of correlation with human judgment needed, and other factors such as computational resources and reference availability. In Table 3, we have compiled a set of metrics. These are categorised into classification, ranking, statistical, model-based, and LLM-based, each with its strengths and weaknesses, discussed below.

Traditional classification metrics They categorise the outputs as correct or incorrect and provide simple performance metrics. The metrics included in this category in Table 3 are accuracy, precision, recall, and F-score. **Strengths**: Traditional classification metrics tend to be simple and intuitive, making them easy to implement and easy to interpret. They are a straightforward tool for binary classification tasks, such as comparing generated text to a gold standard reference. **Weaknesses**: These metrics require a narrow definition of correctness, which is not available in all text generation tasks. They often do not work well with unbalanced datasets, and high-quality labelled data is required to compute these metrics, which is not always available. Furthermore, the binary categorisation approach fails to account for partially correct outputs, which is crucial in open-ended text generation tasks.

Ranking metrics These are used to evaluate the performance of models that produce a ranked set of possible solutions to the given task. These metrics compare these ranked outputs against a correct solution or reference. The metrics included in this category in Table 3 are SaCC (strict accuracy), LaCC (lenient accuracy), and MRR (mean reciprocal rank) (Voorhees and Tice, 2000). **Strengths**: Ranking metrics are easy to understand and allow for more nuanced evaluation of non-deterministic models by considering a ranking of multiple different responses to the same query. They are particularly valuable for recommendation systems (e.g. search engines), where the position of the correct answer in a list is significant (Jadon and Patil, 2025). These metrics offer a more detailed insight into text generation performance than metrics based on only a single output per query.

Weaknesses: The quality of these metrics relies heavily on the references used and can be sensitive to class imbalance. They require multiple outputs per query and ranking, making them more computationally expensive than single-output metrics. Ranking metrics often only consider the highest-ranked correct answer, ignoring other correct answers or when the text generator indicates that the answer is unknown.

Statistical metrics They measure the level of correspondence or matching between n-grams, which can be characters, word pairs or word sequences. The metrics included in this category in Table 3 are chrF (Popović, 2015), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and Perplexity (Jelinek et al., 1977). **Strengths**: They are simple and easy to implement, and especially powerful for NLG tasks that require exact matching, such as speech recognition or machine translation (MT). **Weaknesses**: They do not account for the fact that the same idea can be correctly expressed in various ways, i.e., they do not account for meaning, only for lexical overlap. They are only representative if high-quality references are available.

Model-based metrics These metrics use the tokenizer functions and/or embeddings of language models to encode the text, and evaluate text by computing the similarity between the generated text and the expected answer. The metrics included in this category in Table 3 are BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2019), Mauve (Pillutla et al., 2024), MoverScore (Zhao et al., 2019), COMET (Rei et al., 2020), FrugalScore (Kamal Eddine et al., 2022), and CTRLEval (Ke et al., 2022). **Strengths**: Model-based metrics take semantics into consideration, which traditional metrics cannot do. Whereas traditional metrics tend to only work with proxies for the semantics of a text, the LLMs that these metrics are based on are designed to capture meaning. In some cases, references are not needed. **Weaknesses**: The effectiveness of these metrics depends on the performance of the model. They require significant computational re-

	Metric name	Task	Measures	GT	Resources
Classification	**Accuracy**	-	Fraction of correct answers	✓	- -
	Precision	-	Fraction of true positives over all positive responses	✓	- -
	Recall	-	Fraction of true positives over all actual positive responses	✓	- -
	F-score	-	Harmonic mean of precision and recall	✓	- -
Ranking	**SaCC**	-	How often highest ranked answer is correct	✓	- -
	LaCC	-	How often correct answer in top 5	✓	- -
	MRR	-	How high correct answer ranks on average	✓	- -
Statistical	**chrF**	MT	Character N-gram F-score	✓	-
	BLEU	MT	Similarity of translation to a reference text	✓	-
	ROUGE	Summ.	Lexical overlap between summary and reference text	✓	-
	METEOR	MT	Similarity of translation to a reference text	✓	-
	Perplexity	-	Model's probability to predict a given text	✗	+
Model-based	**BLEURT**	-	Non-trivial semantic similarities between sentences	✓	+
	BERTScore	-	Contextualised embedding-based semantic similarity	✓	+
	Mauve	-	Statistical gap between two text distributions	✓	+
	MoverScore	-	Semantic similarity of generated text to reference text	✓	+
	COMET	MT	Quality of generated translation	✗	++
	FrugalScore	-	Semantic similarity of generated text to reference text	✓	+
	CTRLEval	Gen.	Coherence, consistency and attribute relevance of controlled text generation	✗	+
LLM-based	**G-Eval**	-	Quality of NLG output based on user-defined criteria	✗	+++
	Prometheus 1 and 2	-	Any long-form text based on customised score rubric provided by the user	✓	++
	GEMBA	MT	Quality of generated translation	✗	+++

Table 3: Overview of metrics to evaluate text generation. *GT* = Ground Truth required. For the column *Task* (task specific metrics): *MT* = Machine Translation, *Summ.* = summarization, *Gen.* = text generation.

sources and inherit some of the problems of language models, such as biases. These metrics become a 'moving target' when new and improved versions of the model are released, where the question of whether the metrics should be updated is hard to answer, because although the model might have improved, scores are only comparable between identical models.

Generative LLM as evaluator (LLM-based) These methods leverage the reasoning and instruction-following capabilities of generative LLMs by proposing a tool, framework, or set of steps to use an LLM to evaluate generated text based on some criteria defined by the user. This approach is also often called LLM-as-a-Judge (Gu et al., 2025). The metrics included in this category in Table 3 are G-Eval (Liu et al., 2023a), Prometheus 1 and 2 (Kim et al., 2024a,b), and GEMBA (Kocmi and Federmann, 2023). **Strengths**: These evaluation methods are flexible and can be used with any text generation task. The scores given can be explained by the LLM in natural language. They can be improved by evaluating the same task several times and computing the average or distribution. **Weaknesses**: These metrics are highly dependent on the model, instructions, and language used. They inherit issues like inconsistency, lack of transparency, and bias from language models. Using LLMs as evaluators demands significant computational resources and raises ethical and environmental concerns, especially with proprietary models. Generative LLMs have larger computational requirements than those used in simpler approaches like BERTScore.

4. Evaluation in Practice

To collect practical evaluation challenges, we conducted a survey with 19 LLM developers at a Dutch applied research institute. The survey aimed to gain insights into the evaluation methods currently used and the challenges encountered. Additionally, we implemented metrics in a practical use case to gain insight into their strengths and weaknesses.

4.1. Survey Results Overview

The 19 respondents varied in experience, sector (e.g., government, health), and type of developed applications. We asked the respondents questions on applied evaluation methods, arguments for their choice of evaluation methods, challenges they face in their evaluation, which challenges they find most urgent and what they would need for better evaluations[1]. From the survey results, we focused on respondents' evaluation methods, the reasons for using those, and their questions and needs regarding LLM evaluation in practice. For each response, we manually extracted a list of needs, then grouped similar or identical needs. Finally, we categorised the extracted needs into five themes: Resources, Datasets, Metrics, System, and Governance, which are discussed below[2].

Resources 14 respondents expressed a strong need for comprehensive and practical guidance to support evaluation workflows. This includes general and task-specific evaluation frameworks, prompting strategies, and human-in-the-loop methodologies. There is a clear demand for structured approaches that help practitioners tailor metrics to use cases, integrate manual and automated methods, and identify suitable domain experts. The community also seeks clarity on when and how to apply different evaluation techniques, supported by up-to-date lists of tools and relevant literature. Handling LLM-specific challenges, such as hallucinations, multiple correct answers, and instruction-following failures, was also highlighted as a critical area requiring dedicated guidelines.

Datasets 5 evaluation practitioners emphasised the importance of transparency in dataset construction and usage. Key concerns include contamination of evaluation datasets, the quality and reliability of reference data, and the availability of benchmarks that are both accessible and widely accepted. There is also a need for support in creating custom benchmarks tailored to specific tasks or domains. Notably, respondents highlighted the challenge of evaluating without ground truth, pointing to a gap in methods that can assess model outputs in contexts without a ground truth available.

Metrics 19 practitioners expressed a need for metrics that are interpretable, robust, and aligned with human judgment, especially in terms of correctness over fluency. There is a desire for flexibility in selecting metrics across multiple dimensions, such as factuality, creativity, and bias, and for tools that support custom metric design. Respondents also indicated a desire for visibility into the computational and environmental costs of metrics, as well as guidance on their applicability across different tasks, budgets, and domains. Specific needs include metrics for multilingual evaluation, fuzzy matching, and RAG systems.

System 3 practitioners indicated the importance of contextualising evaluation within the broader system in which the LLM operates. This includes understanding the required quality level for a given application and clarifying who the intended end user of each evaluation method is. There is also a need for clearer guidance on when LLM-based evaluators are appropriate, especially in relation to human judgment and task complexity. These insights point to a growing recognition that evaluation cannot be isolated from the product or pipeline context and must be adapted to real-world deployment scenarios.

Governance The governance theme reflects concerns about the broader implications of evaluation practices. 2 respondents called for consideration of the environmental impact of evaluation methods, as they noted that robust evaluation requires repeated trials to achieve statistically significant results, which can conflict with sustainability goals due to the increased computational and energy demands. There is also a need for scoring or assessing evaluation approaches based on their alignment with ethical and legal standards, particularly EU and Dutch values. These concerns highlight the importance of responsible evaluation practices that go beyond technical performance to include sustainability and compliance dimensions.

4.2. Metric tests in a practical use case

To understand the challenges of evaluating LLM-generated text in practice, we implemented three metrics in a practical use case involving a chatbot prototype for a government dashboard. The chatbot helps users interpret complex health statistics from a psychological/behavioural model. The chatbot is designed to handle simple questions only, allowing psychologists more time to address complex queries from the users of the dashboard. Examples of questions that the chatbot can answer are: *what does 'technostress' mean?*; or *what is the score (or status) of 'social support' in my team*; or *how does 'work-life balance' relate to 'work engagement'?*. The chatbot can also answer more complex filter questions such as: *which indicators in the positive functioning category have a 'green' status?*.

[1] All survey questions can be found in Appendix A.

[2] The full list of needs for LLM evaluation in practice extracted from the survey and categorised by theme can be found in Appendix B.

To answer the questions, the chatbot retrieves data using predefined functions (function calling), which ensures responses are based on actual data. If users ask unrelated questions, the chatbot politely declines to answer. Metrics were chosen based on task type, availability of reference text, and chatbot response type (textual/numerical output). For all tasks, we check that the correct function is being called. Then, by using Table 3 and based on their popularity and availability off-the-shelf, we chose a set of metrics to apply to the content of the responses by the chatbot.

Exact string matching Although not an LLM evaluation metric, in this case, it is suitable for questions in which the chatbot is expected to answer with a numerical score, a predefined status, or a partially fixed message.

For a numerical score or a predefined status (e.g., 'green' or 'good', 'orange' or 'attention', or 'red' or 'urgent'), we built a helper function that extracts the number or status word from the answer and matches it with the expected number or status word from the database.

For queries to which the chatbot refuses to answer, we compiled a set of semantically equivalent formulations in Dutch. In our automatic evaluation, we check that at least one of these pieces of text is present in the chatbot's answer.

This method is simple, reliable and deterministic, and it successfully detects when a chatbot's answer is wrong, according to human judgment.

BLEU (Papineni et al., 2002) For questions about the definition of a term, which involves retrieving definitions from the database, BLEU can measure lexical overlap between the references and the chatbot's output. We used SacreBLEU (Post, 2018), which standardises tokenisation for consistent scores. As the desired output is composed of two different text components (definition and interpretation), we also applied fuzzy string matching with these two texts to complement the BLEU scores. BLEU measures lexical overlap well when small deviations from a reference answer are acceptable. This metric is relatively simple, consistent and deterministic. The score successfully detects when the chatbot response is very different to what is expected. However, because BLEU is a continuous score, it was necessary to define a threshold to classify responses as correct or incorrect. To determine this threshold, we explored three approaches.

The first approach was to use the arithmetic mean of all BLEU scores as the threshold. This method is simple and intuitive, as it considers the overall distribution of scores. However, it is sensitive to outliers, potentially leading to a threshold that does not reflect the majority of cases.

To mitigate the effect of outliers, we also considered the median, which represents the middle value when all scores are sorted. The median is more robust to skewed distributions and extreme values, providing a more stable threshold in cases where the BLEU scores are not symmetrically distributed.

Finally, we applied the Jenks natural breaks optimisation (Fisher-Jenks algorithm) (Jenks, 1967), which partitions the data into classes by minimising variance within each class and maximising variance between classes. In our case, we specified two classes (correct vs. incorrect), and the algorithm identified the break point that best separates the two clusters of BLEU scores. Unlike the mean or median, this method adapts to the actual distribution of the data, making it particularly suitable when scores form distinct groups.

To validate these thresholds, we conducted a manual evaluation comparing expected answers to model outputs. The threshold identified by the Jenks-based approach corresponded best with human judgment. Consequently, we incorporated fixed lower and upper bounds (10 and 30) alongside the Jenks-based threshold for classifying ambiguous responses, ensuring that only responses with a reasonable degree of overlap are considered correct, while also automating the classification of clearly inadequate or highly satisfactory responses.[3]

BERTScore (Zhang et al., 2019) When a user asks for concepts that are not in the database, or about any information unrelated to the dashboard, the chatbot should politely decline to answer. We evaluated this using exact string matching. In addition, we wanted to assess whether a model-based metric would be more suitable than string matching in this case, so we implemented BERTScore. Model-based metrics that calculate the similarities using the contextualised embeddings are more suitable than metrics based on lexical overlap, such as BLEU, for textual outputs in which the response can convey the correct message but using different words. We compared BLEU, fuzzy string matching, and BERTScore on a sample reference text. BERTScore yielded results comparable to BLEU and fuzzy string matching. BERTScore is more lenient towards small syntactic variations, such as synonyms, but the broader context of the sentence is overlooked. The variability in scores is even lower than with BLEU, often lacking a clear cut-off point.

G-Eval (Liu et al., 2023a) In the absence of reference texts, an LLM as evaluator offers flexibility similar to human evaluation, as the model can be provided with specific criteria for assessment. For this reason, we applied G-Eval to complex, open-ended questions in which the user asks the chatbot

[3] A BLEU score of 30 was selected as the upper threshold, as the chatbot responses include additional text beyond the retrieved definition. In our tests, a score of 30 consistently indicated high-quality, complete definitions.

for advice, given the data shown in the dashboard. We used GPT-4o as the backbone LLM for this.

Within this task, the chatbot performs function calling to aggregate data from different parts of the model and provides a recommendation. One of the developers of the chatbot evaluated the chatbot's output according to a list of criteria. We then used the same criteria within G-Eval (Liu et al., 2023a). We focused on the advice part of the response, where the chatbot recommends which scores for terms associated with a broader category should be improved. A requirement was that the chatbot must not generate any information that cannot be derived from the dashboard data. We executed this evaluation repeatedly over the same set of four different responses from the chatbot to the same question. We observed that the scores given by G-Eval are not consistent across the different executions. We also saw that G-Eval did not always comply with the predefined criteria.

5. Lessons Learned

This section summarises the key lessons learned from the literature review, survey, and use case. We share a set of best practices to help developers design effective evaluation strategies for LLM evaluation.

1. The first step for LLM evaluation is to clearly define the evaluation task. This starts with establishing the goal: different objectives require distinct evaluation methods, as we observed in Section 4.2.

2. Identify and describe the end-user of the generated text that is being evaluated.

3. Based on the evaluation description and end-user, choose a suitable evaluation method and metric. When possible, it is advisable to include a form of human evaluation, especially for complex generated answers, or to ensure that the evaluation method performs as expected.

4. Finally, document the choice for the evaluation method and the reasons behind the choice, so that it can be reviewed and adjusted if the evaluation scenario changes in the future. Note that it is also wise to consider intermediate outputs and the number of prompts or the prompting technique needed to reach the final correct answer.

There are several key considerations throughout this process. When using references or a ground truth, these should be of the highest quality possible (i.e., reputable source, verified by domain experts, in line with the defined task). When using non-deterministic metrics (i.e. LLM as evaluator), scores should be computed several times due to potential inconsistencies. Averages and standard deviations should be computed to make the outcomes more robust.

Finally, based on our analysis, survey, and use case, we conclude that human evaluation remains essential for assessing LLM performance. The use case demonstrated that evaluation metrics can be combined and tailored to the specific task or domain. However, even well-designed metrics cannot capture all relevant aspects of model performance. Automated benchmarks have limitations that cannot easily be overcome, making it unwise to rely solely on their outcomes. Instead, meaningful evaluation should combine task-specific metrics with human judgment in real-world contexts, ensuring a more comprehensive and reliable assessment.

6. Conclusion

The evaluation of LLMs is a crucial topic given their rapid development and widespread adoption across domains. As the number of LLM-based applications grows, the need for robust evaluation frameworks becomes evident, not only to guide development and to enable meaningful comparisons between models, but also to ensure their reliability and correctness in critical applications.

In this work, we gathered insights and lessons learned both from scientific literature and from practice. We outlined literature on the evaluation of LLMs and gave a categorisation of the main challenges in LLM evaluation, with a focus on LLM intrinsic, human-related, and automatic evaluation challenges. Additionally, we presented a categorisation of popular and available text generation metrics.

To gain insights into challenges from practice, we conducted a survey targeted to LLM developers. The results show that the main problems mentioned are the absence of references or a ground truth and the difficulty of finding trustworthy benchmarks. According to the respondents, integrating automatic methods with human expertise in evaluation is a relevant direction of future research.

In applying LLM evaluation metrics to a real use case, we found that starting with the simplest metrics that best fit the task and available resources was effective. More sophisticated metrics like BERTScore (Zhang et al., 2019) do not necessarily provide better evaluation quality than simpler ones like BLEU (Papineni et al., 2002) or even string matching, depending on the evaluation's goal. While more flexible metrics like G-Eval (Liu et al., 2023a) can be powerful when no reference text is available, they are difficult to control in terms of consistency and require clearly defined criteria. Ultimately, we found that combining simpler and sophisticated metrics strikes a good balance between evaluation performance, explainability, and resource efficiency. When paired with human evaluation, such a combination can reduce the amount of manual work while still ensuring a proper and

reliable assessment of model performance.

For future work, it would be valuable to test more metrics and combinations of them in a wider array of practical use cases to further highlight evaluation challenges beyond benchmarks and scientific tasks. Expanding metrics to address bias, fairness, and sustainability is critical, given that current frameworks fail to adequately capture these dimensions. Finally, greater transparency from language models about their evaluation processes is needed to facilitate more robust and interpretable assessments.

7. Limitations

This study has several limitations. While we reviewed relevant recent literature, we did not conduct a systematic literature review, so some relevant works may have been missed. The goal of this work was to compare scientific challenges to practical insights, rather than to provide a comprehensive overview. Similarly, while we described and categorised commonly used metrics, the list is not exhaustive.

Another limitation is that our findings and lessons learned are based on a small survey and a single use case, which may not fully capture the broader landscape of LLM evaluation challenges. The survey participants were LLM engineers working in applied research, so our best practices may not fully reflect end-user satisfaction or concerns arising in industry deployment contexts. Involving a broader and more diverse group, such as end users or researchers from different domains, could provide additional perspectives.

Regarding the use case, we tested a limited set of metrics on a single application, which may limit the generalisability of our findings. Due to time constraints, we did not perform a comprehensive statistical analysis of the metrics applied to the different tasks performed by the chatbot, which may affect the reliability and robustness of our observations. Furthermore, due to confidentiality constraints, we are unable to share all details of the use case, which impacts transparency and reproducibility.

Finally, although we touch on broader LLM evaluation challenges, this study primarily focuses on text generation. Evaluating tasks like reasoning or retrieval may require different approaches and further investigation.

8. Ethical Considerations

The survey conducted as part of this study was anonymous, and participants provided informed consent before beginning the survey. No personally identifiable information was collected, and responses were used solely for research purposes. As the survey focused on professional experiences with LLMs and did not involve sensitive personal data, no formal ethics board approval was required.

Acknowledgements

The authors extend their gratitude to the GRAIL project for funding this research. In addition, they would like to thank the participants who filled in the survey, and the use cases for gathering lessons learned in practice.

9. Bibliographical References

Bhashithe Abeysinghe and Ruhan Circi. 2024. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches.

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. Llm stability: A detailed analysis with some surprises.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024.

Lessons from the trenches on reproducible evaluation of language models.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. *arXiv preprint arXiv:2006.14799*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3).

Alejandro de León Languré and Mahdi Zareei. 2024. Improving Text Emotion Detection Through Comprehensive Dataset Quality Analysis. *IEEE Access*, 12:166512–166536. Conference Name: IEEE Access.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jonatas S. Grosman, Pedro H. T. Furtado, Ariane M. B. Rodrigues, Guilherme G. Schardong, Simone D. J. Barbosa, and Hélio C. V. Lopes. 2020. Eras: Improving the quality control in the annotation process for Natural Language Processing tasks. *Information Systems*, 93:101553.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations.

Aryan Jadon and Avinash Patil. 2025. A comprehensive survey of evaluation techniques for recommendation systems. In *Computation of Artificial Intelligence and Machine Learning*, pages 281–304, Cham. Springer Nature Switzerland.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

George F. Jenks. 1967. The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography*, 7:186–190.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Prentice Hall.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning Cheaper, Lighter and Faster Evaluation Metrics for Automatic Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318,

Dublin, Ireland. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An Unsupervised Reference-Free Metric for Evaluating Controlled Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics.

Sunder Ali Khowaja, Parus Khuwaja, Kapal Dev, Weizheng Wang, and Lewis Nkenyereye. 2024. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *Cognitive Computation*, pages 1–23.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative AI: openwashing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1774–1787, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao

Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023c. LLM360: Towards Fully Transparent Open-Source LLMs. ArXiv:2312.06550 [cs].

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Arbi Haza Nasution and Aytuğ Onan. 2024. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, 12:71876–71900. Conference Name: IEEE Access.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A survey on large language model benchmarks.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2024. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024.

How much are llms contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

June Sallou, Thomas Durieux, and Annibale Panichella. 2024. Breaking the silence: the threats of using llms in software engineering. In *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, ICSE-NIER'24, page 102–106, New York, NY, USA. Association for Computing Machinery.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Laurène Vaugrante, Mathias Niepert, and Thilo Hagendorff. 2024. A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions. ArXiv preprint arXiv:2409.20303.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36:77013–77042.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023a. Large language models are diverse role-players for summarization evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 695–707. Springer.

Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy, and Noah A. Smith. 2023b. Transparency Helps Reveal When Language Models Learn Meaning. *Transactions of the Association for Computational Linguistics*, 11:617–634.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. ArXiv:2311.01964 [cs].

Tiffany Zhu, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2024. Human Bias in the Face of AI: The Role of Human Judgement in AI Generated Text Evaluation. *arXiv preprint arXiv:2410.03723*.

A. Appendix A: Survey Questions

Best practices & practical challenges in the evaluation of LLMs As part of our research on the evaluation of Large Language Models (LLMs) we are identifying the challenges with the evaluation of LLMs and the state-of-the-art methods both from scientific and practical perspective. As part of the practical perspective we want to collect both best practices, research and practical challenges developers run into at company XXX. This survey has two goals: 1) Gain insight in how the evaluation is approached in current research on LLMs within company XXX 2) Collect practical challenges that developers/projects run into when evaluating their LLM-based applications The results of this survey will be used anonymised and aggregated (such that they are not traceable to you or your project) in our report and possibly in a scientific paper. By filling this survey you agree with this use of your response. In case of any questions about the survey, please contact XXXX or XXXX Thank you!

General information First we would like to know some general information about your background, experience and sector you are working in.

Q1: What is your background?

- Artificial Intelligence
- Linguistics
- Computer Science
- Mathematics
- Social Sciences
- Other

Q2: What is your experience in working with (Large) Language Models?

- < 1 year
- 1-2 years
- 3-5 years
- 5-10 years
- > 10 years

Q3 For which sector is (most of) your work aimed?

- Health
- Mobility
- Sustainability
- Safety & security
- Government
- Not sector specific
- Other

Your research and applications First we would like to know about your research, in which way do you use LLMs?

Q4: How many projects related to LLMs do/did you work on this year?

- 0
- 1
- 2
- 3
- 4
- 5+

Q5: In what ways do you use LLMs in your research?

- Research on evaluation of LLMs
- Research on increasing performance of LLMs
- Research on integrating LLMs, knowledge and/or tools
- Chatbot
- RAG-application
- Text analysis applications
- Multi-modal applications
- Other

Q6: In which phase is your research? (e.g. problem analysis, design, implementation, evaluation, pilot, deployed) - if you have multiple projects, please answer the question for all of your projects.

Evaluation in practice We would like to hear more about your choices for the evaluation and which challenges you run/ran into?

Q7: Which methods of evaluation do you (aim to) use?

- automatic evaluation (metrics)
- human evaluation
- LLM-based evaluation
- Other

Q8: Can you give a description of the evaluation method you (aim to) use? For example what do you aim to evaluation, which metrics do you use, what do you evaluate with humans?

Q9: Why did you choose this way of evaluation?

Q10: What are challenges that you run/ran into for the evaluation of your LLM?

Q11: What are questions that you have regarding the evaluation of LLMs?

Research on LLM evaluation If you research LLM evaluation specifically, we are curious to hear more about your work.

Q12: Do you research LLM evaluation specifically?

- Yes
- No

Q13: What does your research focus on?

Q14: Do you have results that you can share?

Your opinion on LLM evaluation research? We are curious to hear more about what you think are the most relevant/urgent research topics and challenges.

Q15 What do you think is the most relevant method of evaluation of LLMs? (rank from most to least relevant)

- Human evaluation
- Automatic evaluation
- LLM-based evaluation
- Combination of human and automatic evaluation
- Combination of all three forms of evaluation

Q16: What do you think is the biggest challenge in LLM evaluation (rank form biggest to smalles challenge)

- Robustness, i.e. the ability to produce the similar results for (small) variations in prompts and orders
- Reproducibility, i.e. the ability to produce the same result multiple times
- Data quality, i.e. achieving a qualitative ground truth data set
- Data Leakage, i.e. the fact that test data/ test scenario's might be included in the training data
- Lack of universally accepted benchmarks, i.e. large variety of benchmarks that all have limited variability in prompts/scenario's, are biased to English data, do not consider alternative answers.
- Subjectivity of humans, i.e. different humans give different answers in both annotation and evaluation
- Fairness evaluation, i.e. challenges in gaining insight in biases of the models
- Sustainability, i.e. the energy usage of these models during evaluation and deployment
- Lack of model transparency, i.e. no/limited access or insight in training data and weights
- Other challenge, i.e. a challenge you encounter that is not in this list

Q17: If you ranked 'Other challenge' in the previous question, please let us know which challenge you mean?

Q18: What solution or research would you be helped with?

Next steps This survey is meant as a way to collect more general insights, but does not allow for in-depth discussions. We might be interested to discuss your research and challenges in more detail. These are the last two questions of the survey, thank you very much for helping us and don't forget to submit! :)

Q19: If you are open to a discussion, please leave your email here or send a message to XXX or XXX in case you don't want your answers linked to you.

Q20: If you have any other comments or questions you can leave them here.

B. Appendix B: Survey Results

Resources - Total: 14
15. General guidelines for conducting evaluations.
16. Identification of state-of-the-art evaluation techniques.
20. Methods, frameworks, or guidelines for human/manual evaluation.
23. Guidelines for tailoring metrics to specific use cases.
24. Guidance on integrating task-specific and general evaluation methods.
25. A structured approach to evaluation.
26. Guidance on identifying domain experts for evaluation.
30. Recommendations for effective prompting strategies.
32. Human-in-the-loop evaluation methodologies.
34. An always up-to-date list of available evaluation tools and methods.
35. Guidance on when to use each evaluation method.
36. Clear definitions of text generation tasks to support evaluation.
38. Relevant and rigorous scientific literature on LLMs as evaluators.
43. Guidelines for handling LLM inconsistencies, multiple correct answers, hallucinations, and instruction-following failures.
Datasets - Total: 5
2. Transparency about data contamination in evaluation datasets.
4. An accessible overview of available benchmarks and their quality or acceptance level.
5. Support for creating custom benchmarks.
27. Transparency about the quality of reference data (ground truth).
41. Methods for evaluating without references or ground truth.
Metrics - Total: 19
3. Consistency metrics and repetition guidance tailored to each task.
7. Support for designing custom metrics with heuristic guidance.
8. Specification of the domain or range of applicability for each metric.
9. Visibility into the computational cost of each metric.
10. Clear interpretation of evaluation metrics.
11. Information on the robustness of metrics.
12. Metrics that prioritize correctness over fluency.
13. Frameworks that include multiple evaluation dimensions (e.g., factuality, correctness, conciseness, creativity, bias, interpretability, tone).
14. Flexibility in selecting metrics across different dimensions.
17. Specific metrics for evaluating Retrieval-Augmented Generation (RAG) systems.
18. Indication of how well metrics correlate with human judgment.
19. Guidance on which metrics are suitable under different budget constraints.
22. Mapping between evaluation accuracy (e.g., repetitions needed for significant scores) and environmental cost.
28. Metrics for evaluating the efficiency of LLM responses.
29. Guidance on handling fuzzy matching in evaluation.
31. Indication of speed, accuracy, and completeness of metrics.
39. Capabilities for multilingual evaluation.
40. Indication of which evaluation aspects are prerequisites for others.
42. Insights into the generalizability of metrics.
System - Total: 3
1. A clear understanding of the required quality level for evaluation.
6. Clarity on the intended end user of each metric or evaluation method.
33. Clarity on when LLM-based evaluators are appropriate or not.
Governance - Total: 2
21. Consideration of the environmental impact (e.g., carbon footprint) of metrics.
37. Scoring of evaluation methods based on alignment with EU and Dutch values.

Table 4: Needs for LLM evaluation in practice reported by survey respondents, and grouped under the themes Resources, Datasets, Metrics, System, and Governance.